












An AI agent for treatment reasoning over a biomedical tool universe

Shanghua Gao¹, Ayush Noori^{1,2,3} , Richard Zhu¹ , Curtis Ginder^{1,4} , Zhenglun Kong¹ , Xiaorui Su¹, Justin Kauffman⁵ , Benjamin S. Glicksberg^{5,6,7} , Joshua Lampert^{5,6,8}, Ankit Sakhuja^{5,9,10}, Ashwin Sawant^{5,9,11} , ATHENA-R1 Evaluation Consortium¹², David A. Clifton^{2,13} , Noa Dagan^{3,14,15} , Ran Balicer^{3,14,16} , Marinka Zitnik^{1,3,17,18,19,†} 

¹Department of Biomedical Informatics, Harvard Medical School, Boston, MA

²Department of Engineering Science, University of Oxford, Oxford, UK

³The Ivan and Francesca Berkowitz Family Living Laboratory Collaboration at Harvard Medical School and Clalit Research Institute, Boston, MA, USA

⁴Cardiovascular Division, Department of Medicine, Brigham and Women’s Hospital, Harvard Medical School, Boston, MA

⁵The Windreich Department of Artificial Intelligence and Human Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁶The Hasso Plattner Institute for Digital Health at Mount Sinai, Icahn School of Medicine at Mount Sinai and Mount Sinai Health System, New York City, NY, USA

⁷Mindich Child Health and Development Institute and the Departments of Pediatrics and Genetics & Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁸Mount Sinai Fuster Heart Hospital, Icahn School of Medicine at Mount Sinai, New York, NY, USA

⁹Mount Sinai AI Assurance Lab, Mount Sinai Health System, New York, NY, USA

¹⁰Institute for Critical Care Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

¹¹Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

¹²ATHENA-R1 Evaluation Group (the list of members and their affiliations appears in the Supplementary Information)

¹³Oxford Suzhou Centre for Advanced Research, University of Oxford, Suzhou, Jiangsu, China

¹⁴Clalit Research Institute, Innovation Division, Clalit Health Services, Ramat Gan, Israel

¹⁵Faculty of Computer and Information Science, Ben Gurion University of the Negev, Be’er Sheva, Israel

¹⁶Faculty of Health Sciences, School of Public Health, Ben Gurion University of the Negev, Be’er Sheva, Israel

¹⁷Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University, Cambridge, MA

¹⁸Broad Institute of MIT and Harvard, Cambridge, MA

¹⁹Harvard Data Science Initiative, Cambridge, MA

†Correspondence: marinka@hms.harvard.edu

Abstract

Treatment reasoning underpins every therapeutic decision in medicine, requiring the integration of disease context, comorbidities, concurrent medications, contraindications, and evolving biomedical knowledge to arrive at a therapy appropriate for an individual patient. This process is inherently iterative, as candidate treatments must be evaluated against multiple constraints, revised as new evidence emerges, and grounded in sources that can be inspected and verified. Here we introduce ATHENA-R1, an AI agent for treatment reasoning across all FDA approved drugs since 1939, trained through reinforcement learning over a universe of 212 biomedical tools. At each reasoning step, ATHENA-R1 identifies missing information, selects and executes relevant tools, and incorporates the retrieved evidence before proceeding. To train ATHENA-R1 without human-annotated reasoning traces, we develop a two-level self-learning framework in which multi-agent systems construct tools, treatment tasks, and reasoning trajectories for supervised fine-tuning, followed by reinforcement learning with scientific feedback using rewards for reasoning quality, including evidence gathering, grounded tool use,

and logical non-redundancy, to refine ATHENA-R1’s evidence-seeking strategy. Across five benchmarks spanning 3,168 drug reasoning tasks and 456 patient treatment cases, ATHENA-R1 outperforms language models and tool-use systems. ATHENA-R1 achieves 94.7% accuracy on open-ended drug reasoning and 82.9% accuracy on treatment reasoning, exceeding GPT-5 by 17.8 and 10.7 percentage points, respectively. In blinded evaluations involving experts from 28 rare disease organizations, ATHENA-R1 is preferred over reference models across all evaluation criteria. Physicians rated ATHENA-R1 favorably on complex hospitalized patient cases spanning cardiovascular management and infectious disease. Adverse event hypotheses generated by ATHENA-R1 were tested in electronic health records from 5.4 million patients, with predicted associations reaching adjusted odds ratios of 1.48–1.84 and negative controls showing no elevation. These results establish that treatment reasoning, long considered difficult for AI because it requires knowing what evidence to seek before a conclusion can be formed, can be reframed as a learnable process of iterative evidence gathering, and that reinforcement learning can train AI to perform it.

Main

Treatment reasoning is among the most demanding tasks in medicine, where selecting a therapy for an individual patient requires weighing disease context, patient characteristics, concurrent medications, safety constraints, and evolving evidence [1, 2]. Unlike fact retrieval or pattern recognition, treatment reasoning is an iterative process in which candidate strategies must be gathered, evaluated against multiple constraints, and revised until the evidence supports a decision.

Large language models (LLMs) access medical knowledge through pretraining [3], biomedical alignment [4–7], and agentic frameworks [8, 9]. These models generate fluent responses and capture broad clinical patterns, but they rely on parametric knowledge stored in model weights, lack access to updated and vetted medical information, and can produce recommendations that fail to account for relevant contraindications, interactions or patient-specific constraints. Retrieval augmented generation [10] and tool-augmented LLMs [11–13] can give LLMs access to information outside their model weights, such as medical documents, biomedical databases and software tools, at inference time. However, access to biomedical tools does not by itself produce treatment reasoning. A model must determine what evidence is needed, select the appropriate source, interpret the result in the context of accumulated evidence, and revise its analysis when evidence is incomplete or conflicting. This capacity cannot be assumed from tool access alone and must be learned.

We introduce ATHENA-R1, an AI agent for treatment reasoning that combines multi-step analysis with direct access to medical evidence. Rather than producing answers in a single step, ATHENA-R1 determines what information is needed, retrieves relevant evidence, and uses that evidence to update its analysis. In each reasoning step, ATHENA-R1 selects from a library of 212 biomedical tools, retrieves information about drugs, diseases and patient populations, interprets the

returned evidence, and incorporates it into subsequent reasoning steps. This allows ATHENA-R1 to evaluate candidate treatments through iterative evidence gathering and analysis rather than relying solely on knowledge stored in model weights.

Generating multi-step treatment-reasoning traces at the scale and diversity required for training cannot feasibly be done by human annotators, as each trace must specify what evidence is needed, which tools to call and how retrieved information should be interpreted across hundreds of tools and diverse drug, disease and patient contexts. We therefore train ATHENA-R1 through two sequential stages. First, a multi-agent system automatically constructs biomedical tools, treatment tasks and reasoning traces, yielding ATHENA-R1-INSTRUCT, a dataset of 378,027 instruction-tuning samples derived from 85,340 reasoning traces, comprising 177,626 reasoning steps and 281,695 tool calls grounded in all US FDA approved drugs since 1939. After supervised fine-tuning on ATHENA-R1-INSTRUCT, ATHENA-R1 is refined through reinforcement learning in a live 212-tool environment, receiving rule-based scientific feedback across six dimensions of reasoning quality, including answer correctness, evidence gathering, multi-step reasoning and tool-use validity.

We evaluate ATHENA-R1 on five datasets spanning drug reasoning and patient treatment cases. DrugPC contains 3,168 treatment cases covering 11 treatment tasks, including indications, dosing, safety and pharmacology. BrandPC and GenericPC replace drug names with brand and generic variants, and DescriptionPC replaces names with textual descriptions. TreatmentPC contains 456 treatment cases in which the correct answer depends on patient-specific constraints. Across these datasets, ATHENA-R1 consistently outperforms LLMs and tool-use models in open-ended evaluation. On DrugPC, ATHENA-R1 achieves 94.7% accuracy, exceeding GPT-5 [14] by 17.8 percentage points and DeepSeek-R1 (671B) [15] by 25.9 percentage points. On TreatmentPC, ATHENA-R1 achieves 82.9% accuracy, exceeding GPT-5 by 10.7 percentage points and DeepSeek-R1 by 15.4 percentage points. ATHENA-R1 generalizes across brand names, generic names and diverse drug descriptions (BrandPC, GenericPC and DescriptionPC benchmarks; Extended Data Figure 4; Supplementary Note 1).

We evaluate ATHENA-R1 in three real-world settings. First, experts from 28 rare disease organizations assessed blinded responses to rare disease treatment cases spanning neurodevelopmental disorders, epilepsies, metabolic diseases, rare cancers, channelopathies, and immune-mediated diseases, and preferred ATHENA-R1 over reference models across all eight evaluation criteria, with the largest gains in cognitive traceability and helpfulness of rationale. Second, practicing physicians evaluated ATHENA-R1 on complex hospitalized patient cases in cardiovascular management and infectious disease, including post-CABG patients with CKD, anticoagulated patients with surgical site infections, and post-STEMI patients with severe asthma. Third, we tested adverse event hypotheses generated by ATHENA-R1 in longitudinal health records from 5.4 million patients, prioritizing predictions where prior pharmacovigilance evidence was limited or absent; predicted

associations reached adjusted odds ratios of 1.48–1.84 in the highest risk patient subpopulations, while negative controls remained near null.

Results

ATHENA-R1 reasons over treatment choices by gathering evidence step by step

ATHENA-R1 performs treatment reasoning by combining step-by-step analysis with access to medical evidence (Figure 1). It calls tools from a 212-tool biomedical library (Supplementary Table 3) to retrieve evidence about drugs, diseases and patient populations from curated sources [16, 17]. These tools support queries about indications, contraindications, drug-drug interactions, pharmacology, adverse reactions, disease phenotypes, therapeutic targets and patient-population restrictions. For example, ATHENA-R1 can retrieve a drug’s current approved indications, identify contraindications for a candidate therapy, check interactions between co-medications, map a disease to associated phenotypes, and query target phenotypic evidence. Because these tools are queried in real time, ATHENA-R1 can incorporate current information from FDA prescribing information and biomedical knowledge bases rather than relying solely on knowledge encoded in model parameters. ATHENA-R1 uses the retrieved evidence to guide the next reasoning step, allowing mechanisms, interactions, contraindications and safety constraints to be evaluated together.

At each step, ATHENA-R1 determines what information is needed, selects relevant tools, retrieves evidence and incorporates the returned information into the analysis [18]. It continues this process until the evidence supports a final answer. The output includes both the answer and a reasoning trace that records which evidence was retrieved and how it was used.

ATHENA-R1 can also break complex treatment tasks into smaller analyses. A patient scenario may require identifying candidate treatments, checking drug-drug interactions, evaluating comorbidities, comparing safety warnings and applying patient-specific constraints. ATHENA-R1 analyzes these components and combines the results into a final answer (Figure 1). Additional details of the inference process are provided in Methods 1.2 and Algorithm 1; examples of ATHENA-R1’s key agentic abilities (knowledge grounding, goal-oriented tool selection, multi-step reasoning and real-time retrieval) are shown in Supplementary Figure 1.

Self-learning for treatment reasoning

Multi-step treatment reasoning traces are too large and varied to annotate by hand [19, 20]. Each trace must specify what evidence to retrieve, which tools to use, how to interpret the returned

information and how to combine evidence across multiple reasoning steps. ATHENA-R1 is therefore trained through two levels of self-learning that replace human written traces with generated reasoning trajectories [21]. The first level teaches ATHENA-R1 the structure of treatment reasoning, including problem decomposition, evidence retrieval, tool use and evidence interpretation. The second level teaches ATHENA-R1 how to act within this structure by improving tool selection, evidence gathering and exploration of alternative reasoning paths.

At the first level, ATHENA-R1 automatically constructs its own training data. Generating treatment reasoning traces directly would require a model that already solves the task. Instead, a collection of agent systems generates biomedical tools, treatment tasks and multi-step reasoning traces. This process produces ATHENA-R1-INSTRUCT, a dataset of 378,027 instruction tuning samples derived from 85,340 treatment tasks, comprising 177,626 reasoning steps and 281,695 tool calls grounded in FDA drug labels since 1939 (Extended Data Figure 1). Supervised fine-tuning on ATHENA-R1-INSTRUCT yields the initial ATHENA-R1 model.

At the second level, ATHENA-R1 refines its policy through reinforcement learning (Extended Data Figure 2). During training, ATHENA-R1 explores the 212 biomedical tools used at inference and generates multi-turn reasoning trajectories for each prompt. Each trajectory receives scientific feedback based on rewards for answer correctness [15], output format validity [22], evidence gathering, multi-step reasoning, tool-argument grounding and reasoning non-redundancy. Group relative policy optimization [23, 24] then increases the probability of higher-scoring trajectories. This process improves tool selection and evidence gathering across reasoning steps. Training details are provided in Methods 3.

ATHENA-R1 outperforms language models on Drug Prescribing Card reasoning

We evaluate ATHENA-R1 on DrugPC (Drug Prescribing Cards), a dataset of 3,168 treatment questions (Figure 2a). DrugPC covers 11 tasks of drug information, including drug overview, ingredients, warnings and safety, dependence and abuse, dosage and administration, use in specific populations, pharmacology, clinical information, nonclinical toxicology, patient information and storage and supply. To reduce data leakage from pretraining [25], the evaluation uses drugs approved by the FDA in 2024, while excluding these drugs from training (Methods 4).

We evaluate models in an open-ended setting [5]. Each treatment question is presented without answer choices. The model generates a free-form response, which is then mapped to the correct option among the original 4-5 answer choices [26]. Treatment questions are verified by human experts for validity (Table 2; additional details are provided in Methods 4).

We compare ATHENA-R1 to GPT-5, DeepSeek-R1 and Qwen3 across all 11 tasks (Figure 2b).

ATHENA-R1 achieves 94.7% overall accuracy across the 3,168 questions. GPT-5 achieves 76.9%, DeepSeek-R1 achieves 68.8% and Qwen3 achieves 48.7%. ATHENA-R1 improves accuracy by 17.8 percentage points over GPT-5, 25.9 percentage points over DeepSeek-R1 and 46.0 percentage points over Qwen3. Performance remains consistently high across categories, including warnings and safety, dosage and administration, and use in specific populations. These results show that drug reasoning benefits from iterative evidence gathering and tool use. Rather than relying solely on parametric knowledge, ATHENA-R1 retrieves and interprets FDA label information through multi-step reasoning, improving performance on tasks that require integrating indications, dosing, safety information and population-specific treatment constraints. Ablations isolating the contribution of multi-step reasoning, training trace length, inference step budget, tool execution, and tool library scaling are reported in Extended Data Figure 3 and Supplementary Note 2.

ATHENA-R1 outperforms language models on patient treatment selection

We evaluate ATHENA-R1 on TreatmentPC, a dataset of 456 patient-specific treatment cases. Each case compares drugs indicated for the same disease but differing in properties relevant to treatment selection, including indications, use in specific populations, safety warnings, precautions, contraindications and drug interactions. The correct answer depends on patient context. A treatment may be appropriate for the disease but unsuitable because of pregnancy, comorbidity, dosing constraints or a contraindicated co-medication [27]. We evaluate models in an open-ended setting [5, 26], where the model generates a free form response that is subsequently mapped to one of the answer options (Methods 4).

ATHENA-R1 achieves 82.9% accuracy on TreatmentPC, outperforming all reference models (Figure 2c). ATHENA-R1 exceeds GPT-5 by 10.7 percentage points, DeepSeek-R1 by 15.4 percentage points, Qwen3-Next by 22.8 percentage points and Qwen3 by 43.7 percentage points. Tool use LLMs with access to tool library perform substantially worse. ToolACE-8B [13] achieves 13.4% accuracy and WattTool-8B [12] achieves 5.9%.

These results reveal a central challenge of treatment reasoning: access to biomedical tools alone is insufficient. Models must determine what evidence is needed, select relevant tools, interpret returned information and revise conclusions when evidence is incomplete, conflicting or unexpected.

When GPT-5 is given optional access to tool library, it invokes tools on only 1% of treatment cases and its accuracy falls below its own no-tool baseline. When tool use is required, performance does not recover. In contrast, ATHENA-R1 invokes tools on every treatment case and integrates retrieved evidence into subsequent reasoning steps, achieving 82.9% accuracy. The limiting factor in LLM treatment reasoning is therefore not tool availability but the learned capacity to reason over tool outputs, and providing a frontier model with direct access to tool library does not substitute for

that capacity (Figure 2d; Supplementary Note 2).

TreatmentPC requires joint reasoning over patient case and drug properties. DeepSeek-R1 [15] and GPT-5 [14] are designed for long chain-of-thought reasoning and test-time scaling. To enable multi-step reasoning in DeepSeek-R1, we prompt it using <think> and <\think>. Despite DeepSeek-R1’s 671 billion parameters, ATHENA-R1 outperforms it by 15.4 percentage points (82.9% versus 67.5%). Unlike models that rely primarily on internal knowledge, ATHENA-R1 retrieves FDA labels and drug annotations before answering, allowing each conclusion to be grounded in source evidence. Extended Data Figure 5 illustrates this on a pediatric corticosteroid scenario in which DeepSeek-R1 incorrectly judges the drug as safe, whereas ATHENA-R1 retrieves the FDA label and identifies HPA axis suppression as a documented pediatric risk [28, 29].

Both levels of ATHENA-R1 training contribute to TreatmentPC performance (Figure 2e). In this analysis, the model first generates a free-form reasoning trace and answer, and then selects the answer option that best matches its own response (the “self as judge” protocol; Figure 2d). Under this setting, the Qwen3-8B base model achieves 39.2% accuracy. The first level of ATHENA-R1 self-learning, supervised fine-tuning on ATHENA-R1-INSTRUCT, increases accuracy to 66.5% (+27.3 percentage points). The second level of ATHENA-R1 self learning, reinforcement learning with scientific feedback, further improves accuracy to 74.8% (+8.3 percentage points).

The choice of answer-extraction protocol affects absolute scores but not the direction of the result. Under the “self as judge” protocol, ATHENA-R1 achieves 74.8% accuracy; under the “GPT-5 as judge” protocol used for all baselines, ATHENA-R1 achieves 82.9%. The 8.1 percentage point difference confirms that extraction protocol contributes to absolute performance, but ATHENA-R1’s “self as judge” score of 74.8% still exceeds GPT-5’s 72.2% obtained under the more favorable “GPT-5 as judge” protocol, indicating that ATHENA-R1’s advantage over GPT-5 holds regardless of which answer extraction method is applied.

Disease experts prefer ATHENA-R1 for rare disease treatment reasoning

We evaluated ATHENA-R1 with disease experts from the Chan Zuckerberg Initiative Rare As One network, which brings together a network of rare disease organizations. Rare diseases often lack treatment pathways [30], requiring treatment decisions to be made from limited and fragmented information about disease mechanisms, contraindications, drug interactions and rare disease patient risks [31]. Experts from participating organizations provided disease-specific context used to construct treatment cases spanning neurodevelopmental disorders, epilepsies, metabolic diseases, rare cancers, channelopathies and immune-mediated diseases. The evaluation involved 29 experts from 28 disease organizations (Supplementary Table 4), including clinicians, disease researchers and patient advocates with expertise spanning therapeutic development and clinical care. In total,

23 evaluators completed blinded assessments, contributing 110 expert evaluated responses.

ATHENA-R1 and reference models generated responses with full reasoning traces. The primary reference model was Qwen3-8B, the base model from which ATHENA-R1 is built, chosen to isolate the contribution of ATHENA-R1’s tool library and self learning while holding the underlying language model fixed. Additional comparisons to o3-mini, Gemini-2.0-Flash and DeepSeek-R1 variants (full list in Methods 5) yielded consistent preferences for ATHENA-R1, though the smaller sample per model precludes formal statistical comparison. Experts evaluated paired outputs in a blinded, arena based setting (Figure 3a,b) [5, 32]. For each treatment case, evaluators viewed two responses without model identity and provided pairwise preferences and absolute ratings across eight criteria: task success, helpfulness of rationale, cognitive traceability, possibility of harm, alignment with clinical consensus, accuracy of content, completeness and clinical relevance (Methods 5, Supplementary Note 4) [4, 33]. Pairwise comparisons used four options: “Model A is better,” “Model B is better,” “Both are equally good,” or “Neither did well.” Absolute ratings used a 1–5 Likert scale and an “Unable to Judge” option.

Experts preferred ATHENA-R1 over reference models across all eight criteria (Figure 3c). Percentages were computed over all 110 evaluations; comparison results are reported in Note 4 and Table 3. Preferences were strongest for cognitive traceability (95.5%) and helpfulness of rationale (94.5%), indicating that experts valued responses that exposed the evidence gathering and reasoning process. Across the remaining criteria, ATHENA-R1 was preferred in 57–66% of evaluations, whereas reference models were preferred in 16–20%. Including ties, ATHENA-R1 matched or exceeded reference models in 74–77% of evaluations for completeness (66.4% win, 10.0% tie), task success (63.6% win, 10.9% tie), possibility of harm (61.8% win, 11.8% tie), alignment with clinical consensus (59.1% win, 18.2% tie), accuracy of content (58.2% win, 18.2% tie) and clinical relevance (57.3% win, 18.2% tie).

Absolute ratings confirmed that experts assigned higher quality scores to ATHENA-R1 responses than to reference model responses. ATHENA-R1 achieves a mean score of 4.16 ± 0.90 out of 5, compared with 2.44 ± 1.26 for reference models (Figure 3d). Scores are highest for cognitive traceability (4.67 ± 0.61) and helpfulness of rationale (4.58 ± 0.69), and lowest for possibility of harm (3.76 ± 0.81). The largest gaps occur for cognitive traceability ($\Delta = 3.13$) and helpfulness of rationale ($\Delta = 2.93$), where reference models score 1.54 ± 0.96 and 1.65 ± 1.09 . ATHENA-R1 also improves accuracy of content (3.99 vs. 2.82), completeness (3.88 vs. 2.46) and alignment with clinical consensus (4.18 vs. 2.76). All differences are significant (binomial test and Wilcoxon signed rank test, $P < 5 \times 10^{-5}$; Methods 5, Supplementary Note 4).

ATHENA-R1 reasons through complex, real-world treatment decisions

We next evaluated ATHENA-R1 for patient cases where treatment choices require judgment across incomplete guidelines and competing physiological constraints. We considered five clinical cases from real adult and neonatal hospitalized patients. Three physicians independently rated ATHENA-R1's responses on the three cases selected for formal evaluation, spanning cardiovascular management and infectious disease, using eight criteria on a 1–5 Likert scale: task success, helpfulness of rationale, cognitive traceability, possibility of harm, alignment with clinical consensus, accuracy of content, completeness and clinical relevance. The three formally evaluated cases are shown in Figure 4. Two additional cases, a preoperative polypharmacy case and a neonatal case rated by a single physician, and all five reasoning traces are provided in Supplementary Note 5. Case construction, rubric anchors, ethical framework and statistical treatment are described in Methods 6.

On each case, ATHENA-R1 identified the principal therapeutic risk and proposed a recommendation grounded in cited drug labels, adverse event, and drug mechanism evidence. (1) In a post-CABG patient with CKD stage 2, recent contrast-induced nephropathy, and HFrEF [34], ATHENA-R1 selected enalapril over the pre-admission lisinopril on the basis of reversible effects on BUN and creatinine and the absence of specific CKD contraindications in its label, in contrast to lisinopril's more frequently reported renal dysfunction and hyperkalemia. Reviewer-level means across the eight criteria were 3.50, 3.38, and 3.50, the most consistent agreement of the three cases (standard deviation across reviewers = 0.07; Figure 4a, b; Supplementary Note 5, Case 1). (2) In a patient with a mechanical mitral valve on warfarin who developed a surgical-site infection after total knee arthroplasty with gram-positive cocci on wound culture, ATHENA-R1 flagged the levofloxacin-warfarin interaction [35] as the decisive risk and proposed vancomycin, clindamycin, or linezolid (alongside penicillin-class beta-lactams) as alternatives effective against the cultured organism with more favorable profiles in anticoagulated patients. Reviewer level means were 2.75, 3.63, and 4.25 (SD = 0.75; Figure 4c, d; Supplementary Note 5, Case 2). (3) In a post-STEMI patient with severe persistent asthma and type 2 diabetes, ATHENA-R1 ruled out propranolol on bronchoconstriction grounds, recommended β -blockers with favorable profiles in asthma (metoprolol, bisoprolol, or carvedilol) [36], and flagged hypoglycemia symptom masking as a secondary monitoring concern. Reviewer level means were 4.75, 4.38, and 2.17, the highest overall per-reviewer mean of the three cases but also the widest spread (SD = 1.40; Figure 4e, f; Supplementary Note 5, Case 3), with one reviewer scoring the response markedly below the other two.

Across the three cases, ATHENA-R1 received high ratings for task success, with a mean score of 4.63 ± 0.52 across nine case–reviewer pairs. Every reviewer scored every case at 4 or above for this criterion. Helpfulness of rationale (3.75 ± 0.71) and alignment with clinical consensus were rated above 3 by all reviewers on all cases. Across all 24 case–criterion cells, at least two

of three reviewers rated ATHENA-R1 at 3 or above. Reviewer disagreement was concentrated in cases without a single guideline endorsed treatment, consistent with the intended ambiguity of the selected scenarios. These cases illustrate how ATHENA-R1 reasons through individual treatment decisions rather than providing a systematic quantitative evaluation; the systematic comparison is the cross-organizational expert study reported above. Per-criterion means, standard deviations and handling of reviewer-missing ratings follow the analysis plan in Methods 6.

Population-scale health records support treatment-associated risk hypotheses

We next evaluated whether ATHENA-R1 can generate clinically meaningful hypotheses about treatment-associated adverse event risks. Many adverse events arise not from a disease, comorbidity or medication alone, but from their interaction within specific patient populations. To study this setting, we defined triadic patient cases consisting of a primary disease, a comorbidity and a medication (Figure 5a). For each case, ATHENA-R1 generated candidate adverse events associated with the combined context. To focus on context specific risks, we excluded adverse events predicted from the disease, comorbidity or medication individually and retained those specific to the full triad. Candidate hypotheses were then screened against pharmacovigilance literature by a clinician and a general-purpose LLM [37] and prioritized for analysis where supporting evidence was limited or absent from the prior literature, making EHR evaluations a test of hypothesis generation rather than confirmation of established drug safety signals (Methods Section 7, Supplementary Note 6).

We evaluated these hypotheses using cohort analyses [38] in longitudinal health records from 5.4 million patients (Figure 5b; Methods Section 7). For each predicted adverse event, we measured prevalence and estimated confounder adjusted odds ratios using multivariate regression. Models adjusted for age, sex, socioeconomic status and outpatient healthcare utilization. Residual confounding may remain because clinical factors, treatment history and disease severity are not fully captured by these variables [39] (Supplementary Note 7). Across disease–comorbidity–drug contexts, adverse events predicted by ATHENA-R1 showed higher prevalence in the most specific patient subpopulations and elevated adjusted odds ratios relative to broader comparison cohorts.

Cohort analyses supported ATHENA-R1’s prediction of increased risk of acute kidney failure in patients with hypertension and gout treated with β -blockers (OR = 1.84, 95% CI: 1.69–2.00; Figure 5c, 6a). Antihypertensive therapy has been associated with acute kidney injury (AKI), although prior analyses have not consistently identified a specific signal for β -blockers [40]. β -blockers are widely used in patients with hypertension and chronic kidney disease (CKD) and are generally considered hemodynamically neutral with respect to renal perfusion and glomerular filtration [41]. However, β -blockers increase serum uric acid levels and are associated with incident gout [42]. Hyperuricemia is associated with renal microvascular dysfunction, inflammation and

increased risk of AKI [43]. This supports a pathway in which elevated uric acid levels in patients with gout act as an effect modifier, identifying a subgroup in which β -blocker associated metabolic effects may contribute to adverse renal outcomes. Disease severity and baseline renal function are the variables most likely to contribute to residual confounding in this analysis, as both influence β -blocker selection and independently predict renal outcomes. Clinically, AKI, including severe forms extending to acute renal failure, carries substantial morbidity and mortality, particularly in patients with multiple comorbidities.

For the same patient subpopulation, ATHENA-R1 predicted increased risk of hyperkalemia, supported by cohort analyses (OR = 1.78, 95% CI: 1.59–2.00; Figure 5d, 6a). β -blockers have been associated with a modest but statistically significant increase in hyperkalemia risk after adjustment for comorbidities and monitoring frequency [44]. Mechanistically, β_2 -adrenergic blockade reduces cellular potassium uptake, impairing extrarenal potassium handling and increasing serum potassium levels [45]. As described above, β -blockers increase uric acid levels and gout risk, which is associated with AKI. Because renal excretion accounts for most potassium elimination, reduced kidney function further increases hyperkalemia risk. This defines a sequential pathway from β -blockade to hyperuricemia to AKI to impaired potassium clearance. Clinically, hyperkalemia is a life threatening electrolyte disturbance associated with arrhythmias, sudden cardiac death and interruption of guideline-directed therapies, making identification of high risk subgroups important for monitoring and prevention.

Cohort analyses supported ATHENA-R1's prediction of increased risk of hepatocellular carcinoma in patients with diabetes and ischemic heart disease treated with DPP-4 inhibitors (OR = 1.48, 95% CI: 1.17–1.88; Figure 5e, 6a). Diabetes is a well-established risk factor for hepatocellular carcinoma (HCC), driven by insulin resistance and chronic inflammation leading to steatohepatitis and cirrhosis [46]. DPP-4 inhibitors are oral anti-hyperglycemic agents, although their use has declined relative to SGLT-2 inhibitors and GLP-1 receptor agonists [47]. Evidence linking DPP-4 inhibitors to HCC remains inconsistent, with randomized trial meta-analyses showing no clear increase in risk compared to alternative therapies [48]. By contrast, GLP-1 receptor agonists and SGLT-2 inhibitors show signals of reduced HCC risk [49]. DPP-4 is involved in immune regulation and tumor biology, providing a basis for context dependent effects [50]. Ischemic heart disease likely identifies a population with greater metabolic burden and comorbidity rather than a direct interaction, although this association warrants further study.

Prevalence analyses supported ATHENA-R1's prediction of increased risk of squamous cell carcinoma (SCC) in patients with hypertension and gout treated with diuretics, although adjusted regression did not reach statistical significance (OR = 1.08, 95% CI: 0.66–1.78; Figure 5f, 6a). Thiazide diuretics, particularly hydrochlorothiazide, are associated with increased SCC risk in a dose-dependent manner [51]. The mechanism likely involves photosensitization, which increases

susceptibility to ultraviolet-induced DNA damage. Diuretics also increase uric acid levels and gout risk, and gout has been linked to increased cancer risk [52]. These effects combine to produce a layered risk structure. Clinically, SCC can be locally invasive or metastatic, particularly in older patients with comorbidities. The absence of a significant adjusted odds ratio does not rule out a true association; SCC is a relatively rare outcome in observational data and the analysis is likely underpowered to detect modest effects. This result illustrates the range of confidence that should be applied to model generated hypotheses and motivates prospective validation in larger cohorts.

Prevalence analyses supported ATHENA-R1's prediction of increased risk of liver failure in patients with hyperlipidemia and hypothyroidism treated with statins, although adjusted regression did not reach statistical significance (OR = 1.04, 95% CI: 0.68–1.58; Figure 5g, 6a). Statins can cause liver enzyme elevations, but true liver failure is rare, with rates below 2 per 1 million patient years [53]. Mechanistically, statins can induce hepatocellular injury through mitochondrial and metabolic effects. Hypothyroidism is associated with steatotic liver disease and cirrhosis, which may increase baseline hepatic vulnerability. This suggests a context in which statin exposure contributes to more severe hepatic outcomes in a susceptible subgroup.

ATHENA-R1 predicted increased risk of respiratory failure in patients with diabetes and chronic kidney disease treated with metformin; cohort analyses did not show a significant increase (OR = 1.00, 95% CI: 0.92–1.07; Figure 5h, 6a). Metformin is associated with lactic acidosis in patients with impaired renal function. Reduced clearance increases the risk of drug accumulation and acidosis. Metabolic acidosis increases respiratory drive and can contribute to respiratory decompensation [54]. Observational studies also report increased respiratory complications in high-risk populations with COPD. These mechanisms support indirect pathways linking metformin to respiratory outcomes in specific patient groups.

We validate the analysis using positive and negative controls. Positive controls recover established clinical effects, including hyperkalemia from ACE inhibitor therapy in CKD [55], ischemic risk from inhaled β -agonists [56], and acidosis from metformin in CKD [57]. We also recover the cardiovascular benefits of GLP-1RA and SGLT-2 therapies [58–60] (Figure 6b). Negative controls show no increase in adjusted risk, with estimates centered near null (Figure 6c). These results indicate that the evaluation approach does not generate spurious associations in the absence of plausible mechanisms and supports its use for evaluating plausibility of clinically meaningful adverse event risk.

Discussion

ATHENA-R1 is trained to perform treatment reasoning by gathering evidence through tool use, interpreting returned information, and revising its analysis over multiple steps. This process produces both a treatment recommendation and a reasoning trace that records how evidence contributes to the final conclusion. ATHENA-R1 addresses treatment reasoning tasks in which a treatment case is posed and evidence must be gathered before a conclusion can be reached, tasks that arise across drug selection, personalized treatment planning, and hypothesis generation about treatment-associated risks. It is not a point-of-care reference tool or a risk calculator, but a research system to evaluate whether AI can learn the evidence seeking process that underlies treatment decisions. Across benchmark datasets, expert evaluations, clinical cases, and population-scale analyses, ATHENA-R1 outperformed language models and tool-use systems. Benchmark evaluations establish performance under controlled conditions, while expert evaluations in rare disease settings, physician ratings on complex hospitalized patient cases, and population-scale EHR validation test whether that performance extends to open-ended cases with real-world clinical complexity.

A central design choice in ATHENA-R1 is the separation of reasoning from knowledge storage. Rather than relying solely on knowledge encoded in model parameters, ATHENA-R1 retrieves evidence from tools that query biomedical resources, including FDA drug labels and clinically validated knowledge bases, allowing conclusions to be grounded in retrievable evidence and updated as new information becomes available. This separation also has implications for how model scale relates to reasoning performance. Consistent performance of ATHENA-R1 above DeepSeek-R1 (671B parameters) and GPT-5 suggests that targeted tool-use training may be more effective than scaling alone for tasks that require iterative evidence gathering. Together, these results show that treatment reasoning can be framed as an iterative process of evidence gathering and analysis, and that reinforcement learning can train AI to perform this process.

Several limitations warrant further investigation. The quality of ATHENA-R1's outputs depends on the coverage and reliability of tool library. Missing endpoints and retrieval errors can propagate through treatment reasoning [61]. ATHENA-R1 also does not quantify uncertainty. Although grounding in retrieved evidence improves verifiability (Extended Data Figure 5), uncertainty can arise from incomplete evidence, incorrect tool outputs, and errors in evidence synthesis [62]. In the clinical case evaluations, reviewer disagreement was highest on cases without clear guideline endorsed answers, precisely the cases where quantified uncertainty would be most valuable [21]. In addition, training relies on generated reasoning traces rather than human written demonstrations. This enables supervision at a scale that would otherwise be impractical, but generated traces may inherit biases from the generation process [63].

ATHENA-R1 operates on natural language inputs and does not incorporate imaging, laboratory

time series, genomic measurements, or longitudinal patient records. Extending treatment reasoning to multimodal patient data will require new representations and evaluation frameworks [64]. Population-scale analyses presented here to evaluate hypotheses generated by ATHENA-R1 are observational. Although these analyses adjust for patient demographics and healthcare utilization (Supplementary Note 7), residual confounding and measurement error may remain. These results should be interpreted as support for hypothesis generation rather than causal inference.

ATHENA-R1 frames treatment reasoning as an evidence seeking process in which AI gathers and evaluates biomedical information before reaching a conclusion. By separating treatment reasoning from knowledge storage, ATHENA-R1 can incorporate new evidence while preserving a transparent record of how conclusions are formed. Results presented here establish a principle that treatment reasoning, which requires knowing what evidence to seek before a recommendation can be formed, can be learned through reinforcement learning over a universe of biomedical tools. As biomedical knowledge expands to encompass genomic, imaging, and longitudinal patient data, and as tool libraries expand to query these sources, training frameworks of this kind may support reasoning across an increasingly broad range of therapeutic contexts.

Acknowledgements. We gratefully acknowledge the support of NSF CAREER 2339524, ARPA-H Biomedical Data Fabric (BDF) Toolbox Program, Harvard Data Science Initiative, Amazon Faculty Research, Google Research Scholar Program, AstraZeneca Research, Roche Alliance with Distinguished Scientists (ROADS) Program, Sanofi iDEA-iTECH Award, GlaxoSmithKline Award, Boehringer Ingelheim Award, Merck Award, Optum AI Research Collaboration Award, Pfizer Research, Gates Foundation (INV-079038), Chan Zuckerberg Initiative, John and Virginia Kaneb Fellowship at Harvard Medical School, Biswas Computational Biology Initiative in partnership with the Milken Institute, Collaborative Center for XDP at Massachusetts General Hospital, Harvard Medical School Dean's Innovation Fund for the Use of Artificial Intelligence, and the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. A.N. was supported by the Rhodes Scholarship. D.A.C. was funded by an NIHR Research Professorship (NIHR302440), a Royal Academy of Engineering Research Chair, and the InnoHK Hong Kong Centre for Cerebro-Cardiovascular Engineering, and was supported by the National Institute for Health Research Oxford Biomedical Research Centre and the Pandemic Sciences Institute at the University of Oxford. This research was enabled by the AI Cluster at the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders. Figures 1, 2, 3, and 5 were created, in part, using Biorender.com (see <https://biorender.com/zn1rz6u>).

Ethics approval. All parts of this study that relate to the use of Clalit Health Services data were approved by the Clalit Health Services Institutional Review Board (Helsinki) committee. Work conducted at Mount Sinai was approved by the Mount Sinai Institutional Review Board (STUDY-20-00338).

Data availability. ATHENA-R1 is available at <https://athena.openscientist.ai>.

Code availability. All code used in this study is available at <https://github.com/mims-harvard/ATHENA> under MIT license. Model weights are released at <https://huggingface.co/mims-harvard/ATHENA-R1-Qwen3-8B>.

Author contributions. S.G. developed and implemented ATHENA-R1, and performed all benchmarking analyses. A.N., J.K., B.G., A.S. and J.L. developed and evaluated the clinical case scenarios. S.G. and R.Z. developed the rare disease treatment reasoning evaluation. Members of the ATHENA-R1 Evaluation Consortium performed human evaluations in rare disease treatment reasoning. A.N., N.D. and R.B. designed and performed the patient cohort analyses using electronic health records. All authors discussed the results and contributed to the manuscript. S.G. and M.Z. designed the study, and M.Z. supervised and led the overall study.

Competing interests. The authors declare no competing interests.

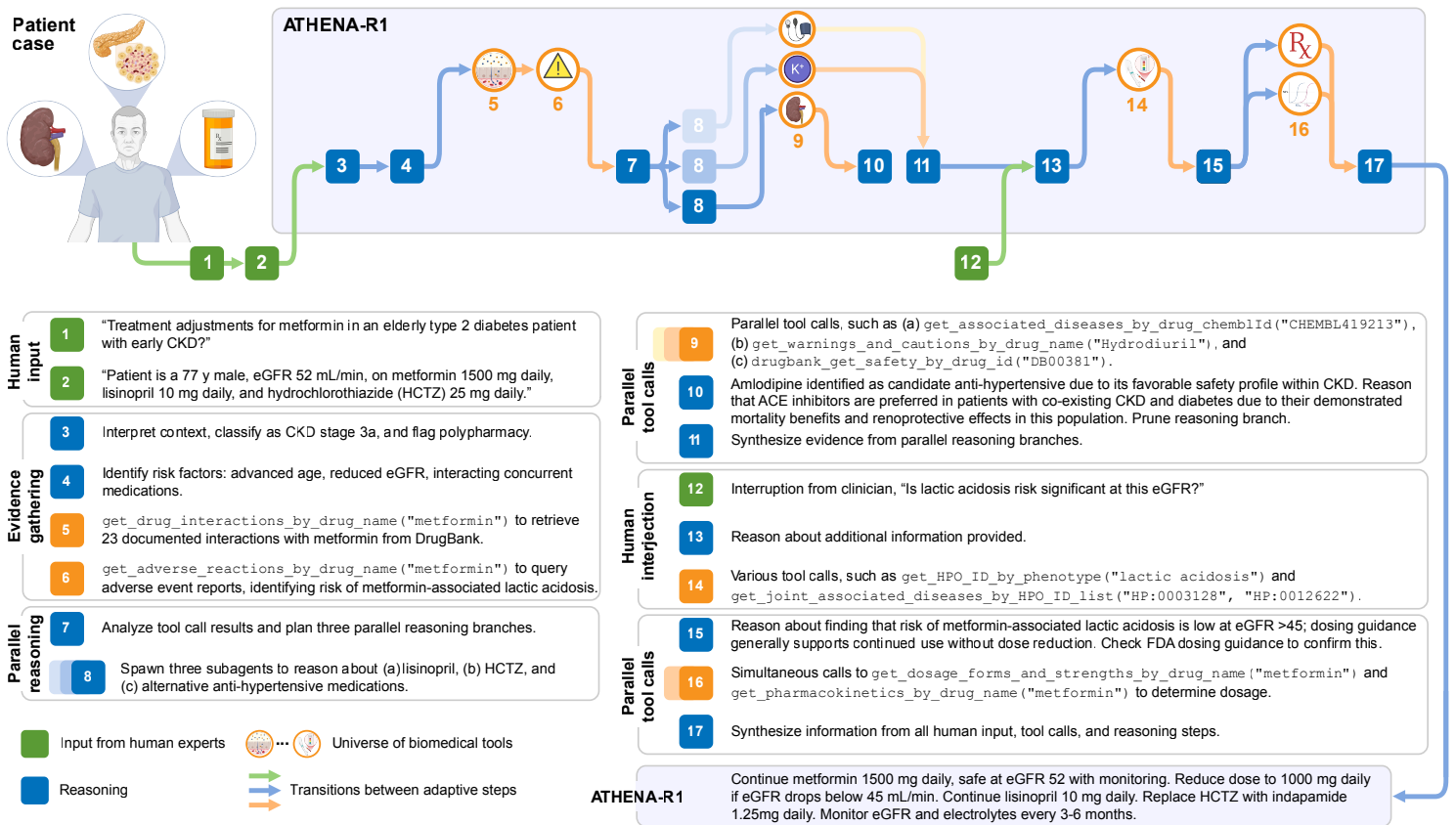


Figure 1: ATHENA-R1 solves precision treatment reasoning problems by retrieving and analyzing medical evidence from a biomedical tool universe. For a patient treatment scenario, ATHENA-R1 generates a treatment recommendation together with a reasoning trace that records evidence retrieval, tool use, and intermediate analyses. The example shows ATHENA-R1 adjusting therapy for a 77-year-old man with type 2 diabetes and early chronic kidney disease (eGFR 52 mL/min) receiving metformin, an ACE inhibitor, and hydrochlorothiazide. Green nodes denote human input and clinician interjections, blue nodes reasoning steps, and orange nodes tool calls that retrieve biomedical evidence. ATHENA-R1 does not follow a fixed sequence of operations. It adaptively determines which evidence to collect, which analyses to perform, and which questions require further investigation based on information gathered in previous steps. In this example, ATHENA-R1 evaluates the ACE inhibitor, hydrochlorothiazide, and alternative treatment options in parallel, eliminates unsupported hypotheses (here, calcium-channel blockers show no interaction with metformin), and synthesizes the remaining evidence. A clinician interjection ("Is lactic acidosis risk significant at this eGFR?") triggers additional evidence gathering and analysis before ATHENA-R1 generates its final recommendation. Multiple orange nodes branching from a single blue node indicate parallel tool calls within one reasoning step. Separate reasoning branches indicate concurrent analyses of different treatment considerations. This adaptive process allows ATHENA-R1 to integrate patient characteristics, medications, contraindications, and biomedical evidence when evaluating treatment options.

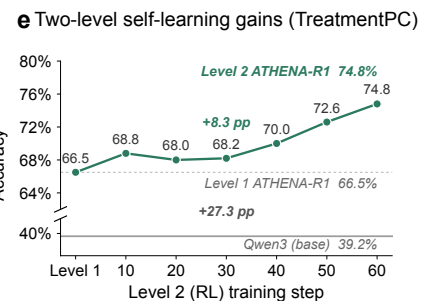
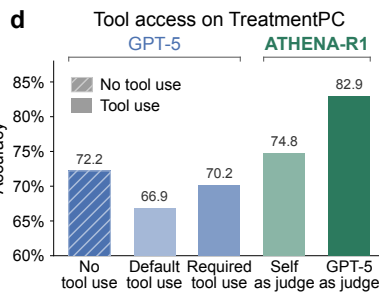
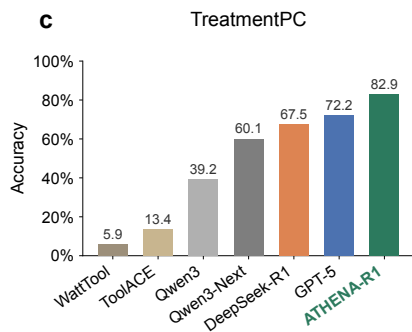
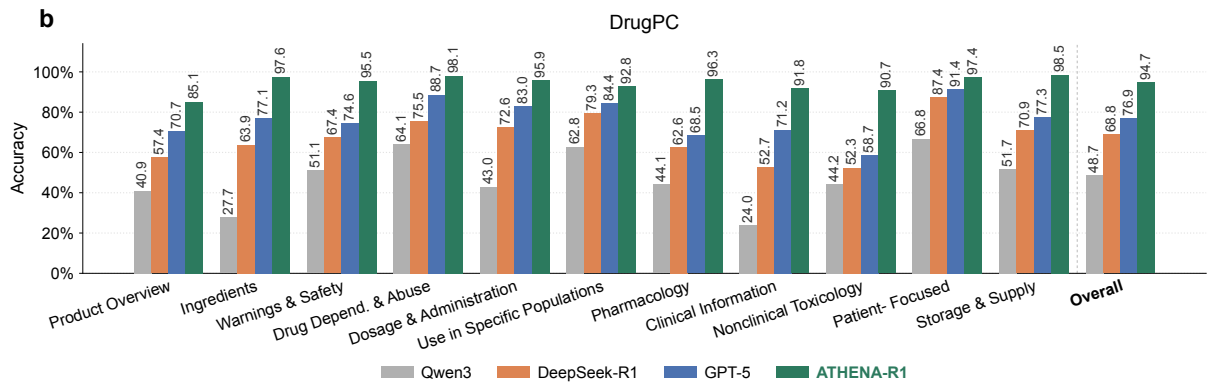
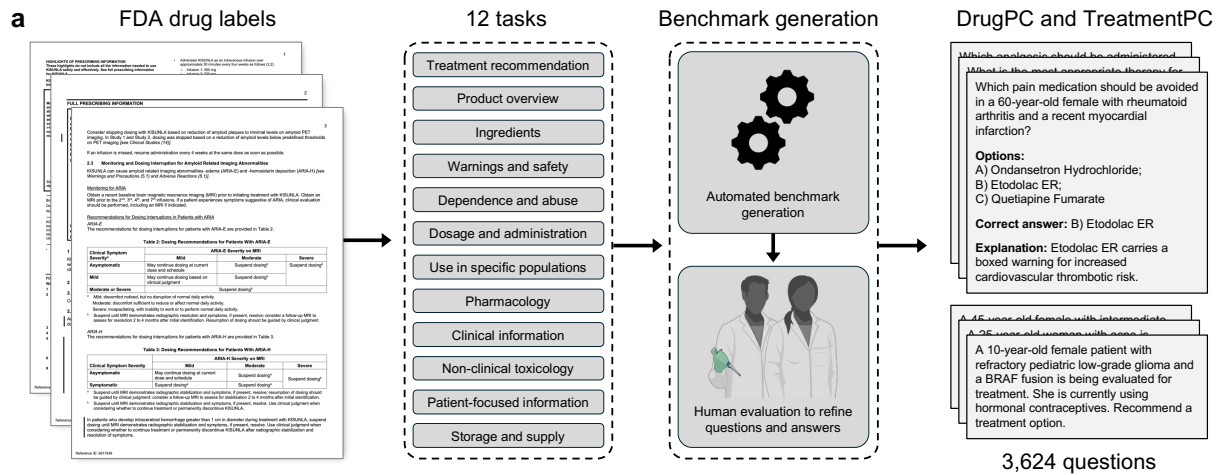


Figure 2: ATHENA-R1 outperforms reasoning models and tool-use LLMs on drug prescribing and patient treatment benchmarks. (a) Construction of the DrugPC and TreatmentPC benchmarks from FDA prescribing information. Structured FDA drug labels were used to generate treatment questions across drug prescribing and patient-specific treatment selection tasks. Human review was used to refine the questions, answer choices, and explanations.

(b) Accuracy on the DrugPC open-ended benchmark, comprising 3,168 questions across 11 drug-prescribing tasks. Each model generated a free-form answer without access to answer options; an independent GPT-5 instance then mapped the answer to one option for scoring, using the protocol shown in **d**. ATHENA-R1 achieved 94.7% micro-averaged accuracy across all questions, compared with 76.9% for GPT-5, 68.8% for DeepSeek-R1, and 48.7% for Qwen3. The 11 tasks cover drug overview, ingredients, warnings and safety, dependence and abuse, dosage and administration, use in specific populations, pharmacology, clinical information, nonclinical toxicology, patient-focused information, and storage and supply. **(c)** Accuracy on the TreatmentPC open-ended benchmark, comprising 456 patient-specific treatment scenarios. ATHENA-R1 achieved 82.9% accuracy, outperforming reasoning models GPT-5 (72.2%), DeepSeek-R1 (67.5%), Qwen3-Next (60.1%), and Qwen3 (39.2%), as well as tool-use LLMs with full access to tool library, including ToolACE-8B (13.4%) and WattTool-8B (5.9%). **(d)** TreatmentPC accuracy under matched tool access and answerextraction conditions. GPT-5 reached 72.2% without tools, 66.9% when tools were available but optional, and 70.2% when required to call a tool on every question. Under optional tool access, GPT-5 used tools on 1% of questions, whereas ATHENA-R1 used tools on every question. Because open-ended answers must be mapped to answer options before scoring, we compared two answer-extraction protocols for ATHENA-R1. In the ‘self as judge’ protocol, ATHENA-R1 mapped its own free-form answer to an option and reached 74.8% accuracy. In the ‘GPT-5 as judge’ protocol, an independent GPT-5 instance performed the same mapping and ATHENA-R1 reached 82.9% accuracy, the protocol used for all baselines in **c**. Both protocols score the selected option against the same gold answer key; they differ only in how answer extraction is performed. **(e)** TreatmentPC accuracy across the two levels of ATHENA-R1 self-learning, scored with the ‘self as judge’ protocol from **d**. The Qwen3-8B base model achieved 39.2% accuracy. Level 1, supervised fine-tuning on ATHENA-R1-INSTRUCT, increased accuracy to 66.5%. Level 2, reinforcement learning with scientific feedback, further improved accuracy across RL steps 0-60, reaching 74.8% at step 60. The y-axis is broken to show the Qwen3 baseline and the RL trajectory on the same plot. Each ATHENA-R1 accuracy in **b-e** comes from one independent rollout; sampling variability across $n=5$ independent rollouts is reported in Extended Data Figure 6 and Supplementary Note 8.

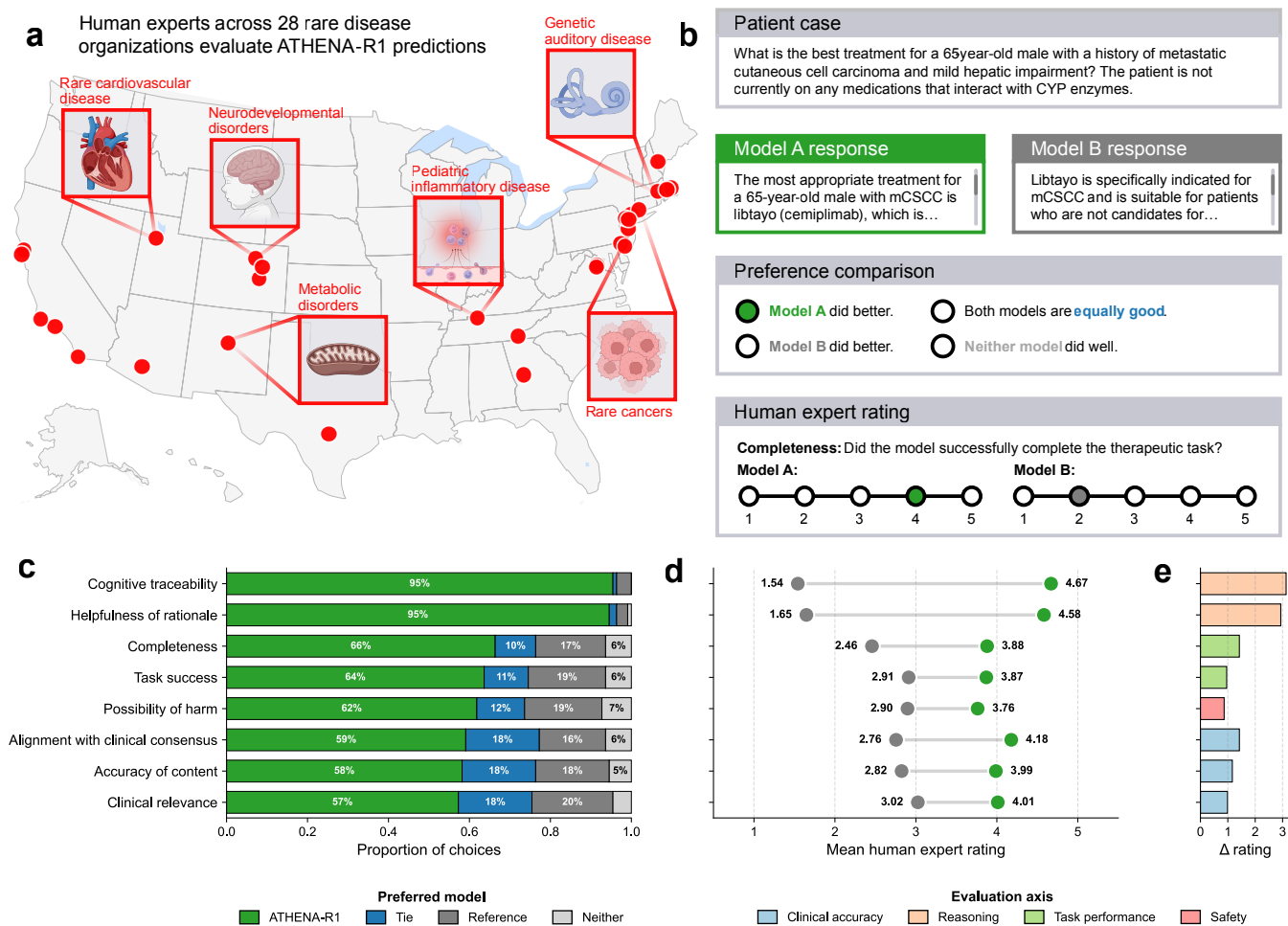
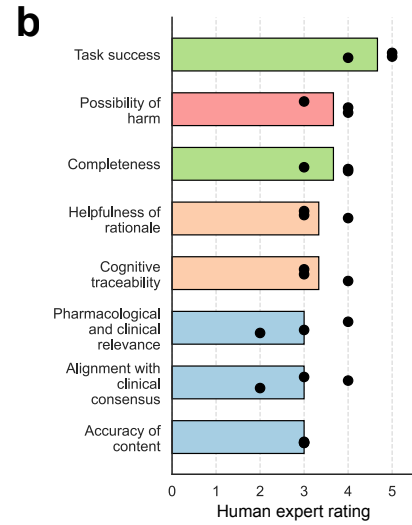


Figure 3: Across all eight evaluation criteria, disease experts from 28 rare disease organizations prefer ATHENA-R1's responses and reasoning traces over those of reference models. (a) Design of the human evaluation of ATHENA-R1. For each patient case and treatment-development scenario, ATHENA-R1 and reference models (predominantly Qwen3-8B; six reference models in total, listed in Online Methods) independently generated a response and multi-step reasoning trace. Each case was routed to experts whose disease expertise matched the case. We collected 110 expert-evaluated responses from 23 evaluators, including disease experts from 28 rare disease organizations in the Chan Zuckerberg Initiative Rare As One network. **(b)** Arena-based evaluation interface. For each treatment case, experts view two responses side by side with model identity hidden, select the preferred response (pairwise preference), and rate each response on a 1–5 scale. Both judgments span eight criteria: task success, helpfulness of rationale, cognitive traceability, possibility of harm, alignment with clinical consensus, accuracy of content, completeness, and clinical relevance. **(c)** Pairwise preferences across all eight criteria. In blinded head-to-head comparisons, experts prefer ATHENA-R1 over reference models on every criterion, with the largest margins for cognitive traceability (95.5%) and helpfulness of rationale (94.5%). **(d)** Absolute ratings across all eight criteria. Experts rate ATHENA-R1 outputs at a mean of 4.16 ± 0.90 out of 5, versus 2.44 ± 1.26 for reference models. **(e)** Δ rating between ATHENA-R1 and the reference model. Bars are colored by evaluation axis.

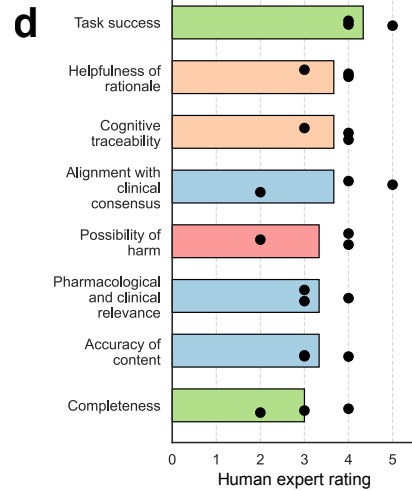
a

Patient 67M POD: #2 s/p three-vessel CABG.	Recent course of disease Patient underwent coronary angiography 5 days pre-op with 75mL low-osmolar contrast, which revealed severe 3-vessel disease. Developed mild contrast-induced nephropathy with Cr rising to 1.7 (eGFR 42). Underwent successful CABG (LIMA->LAD, SVG->OM, SVG->PDA) with 110 minutes on bypass. Initial post-op course required low-dose pressors for 12 hours, now resolved. Current Cr stable at 1.7, good UOP.
Reason for consult ACE inhibitor management in setting of recent contrast exposure.	
Past medical history HTN (15 years), T2DM (HbA1c 6.8%), hyperlipidemia, CKD stage 2 (baseline Cr 1.4, eGFR 52), HFrEF (LVEF 35% per recent echo), myocardial infarction (4 months ago).	
Medication history Lisinopril 20mg daily, metformin 1000mg BID, atorvastatin 40mg daily, ASA 81mg daily.	
	Question <i>What ACE inhibitors can be given to this patient postoperatively?</i>



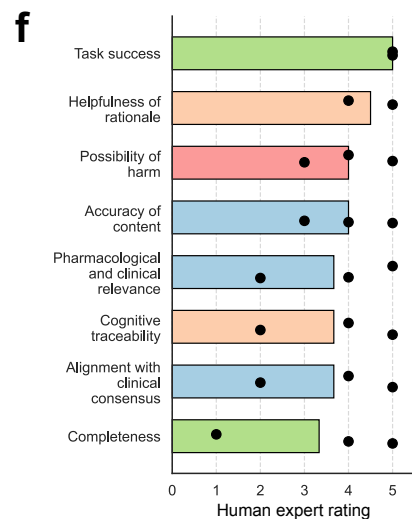
c

Patient 68F POD #4 s/p right total knee arthroplasty.	Recent course of disease Developed SSI after TKA (fever 38.9°C, WBC 14.2k, wound culture gram-positive cocci). Warfarin held peri-op; currently on prophylactic enoxaparin (subtherapeutic for valve) and empirical cefazolin.
Reason for consult Antibiotic selection; ID recommends switching to levofloxacin.	
Past medical history Metallic MVR (2019), chronic AF, HFpEF, CKD stage 3, osteoarthritis.	
Medication history Warfarin (INR 2.5-3.5), metoprolol, furosemide, lisinopril, atorvastatin.	
	Question What are the risks of levofloxacin and are there alternatives?



e

Patient 51F POD #1 s/p PCI for STEMI.	Recent course of disease Admitted with STEMI, underwent successful PCI with drug-eluting stent. Post-procedure stable. Cardiology recommends initiating beta-blocker.
Reason for consult Beta-blocker selection for secondary prevention.	
Past medical history Severe persistent asthma (well-controlled on high-dose fluticasone/salmeterol + albuterol PRN), T2DM (HbA1c 7.2%, on Metformin), HTN, hyperlipidemia.	
Medication history Aspirin 81mg, Clopidogrel 75mg, Atorvastatin 80mg, Lisinopril 5mg.	
	Question <i>Which beta-blocker is most appropriate?</i>



Clinical accuracy
 Reasoning
 Task performance
 Safety

Figure 4: On complex, real-world hospitalized-patient cases in cardiovascular management and infectious disease, physician reviewers consistently rate ATHENA-R1's treatment recommendations as successful.

Each case is a real hospitalized patient whose treatment decision requires weighing competing physiological constraints under incomplete guideline coverage. In each case, ATHENA-R1 identifies the principal therapeutic risk and proposes a recommendation grounded in cited label, adverse-event, and mechanism evidence. Panels pair each case (left) with its expert ratings (right). **(a)** A 67-year-old male on postoperative day 2 following three-vessel CABG, with HFrEF, CKD stage 2 and recent contrast-induced nephropathy, asking which ACE inhibitors can be safely administered postoperatively. **(b)** Expert absolute ratings (1–5 scale) on the case in **a** across eight criteria: task success, possibility of harm, completeness, helpfulness of rationale, cognitive traceability, clinical relevance, alignment with clinical consensus, and accuracy of content. Bars show the mean across three physician reviewers; dots show individual reviewer scores. **(c)** A 68-year-old woman with a mechanical mitral valve on warfarin, 4 days after elective total knee arthroplasty, developed a surgical-site infection with gram-positive cocci on wound culture; the infectious disease team recommended switching to levofloxacin. Question: what are the risks of levofloxacin and what are the alternative antibiotics? **(d)** Expert absolute ratings on the case in **c**, across the same eight criteria; bars and dots as in **b**. **(e)** A 51-year-old female on postoperative day 1 after PCI for STEMI with severe persistent asthma and type 2 diabetes, asking which β -blocker is most appropriate for secondary prevention. **(f)** Expert absolute ratings on the case in **e**, across the same eight criteria; bars and dots as in **b**. Task success received a mean rating of 4.63 ± 0.52 across the nine case-reviewer pairs, and across all three cases and all eight criteria at least two of three reviewers rated ATHENA-R1's response at 3 or above. These cases illustrate ATHENA-R1's reasoning on individual treatment decisions and complement the systematic expert evaluation in Figure 3. Bars are colored by evaluation axis.

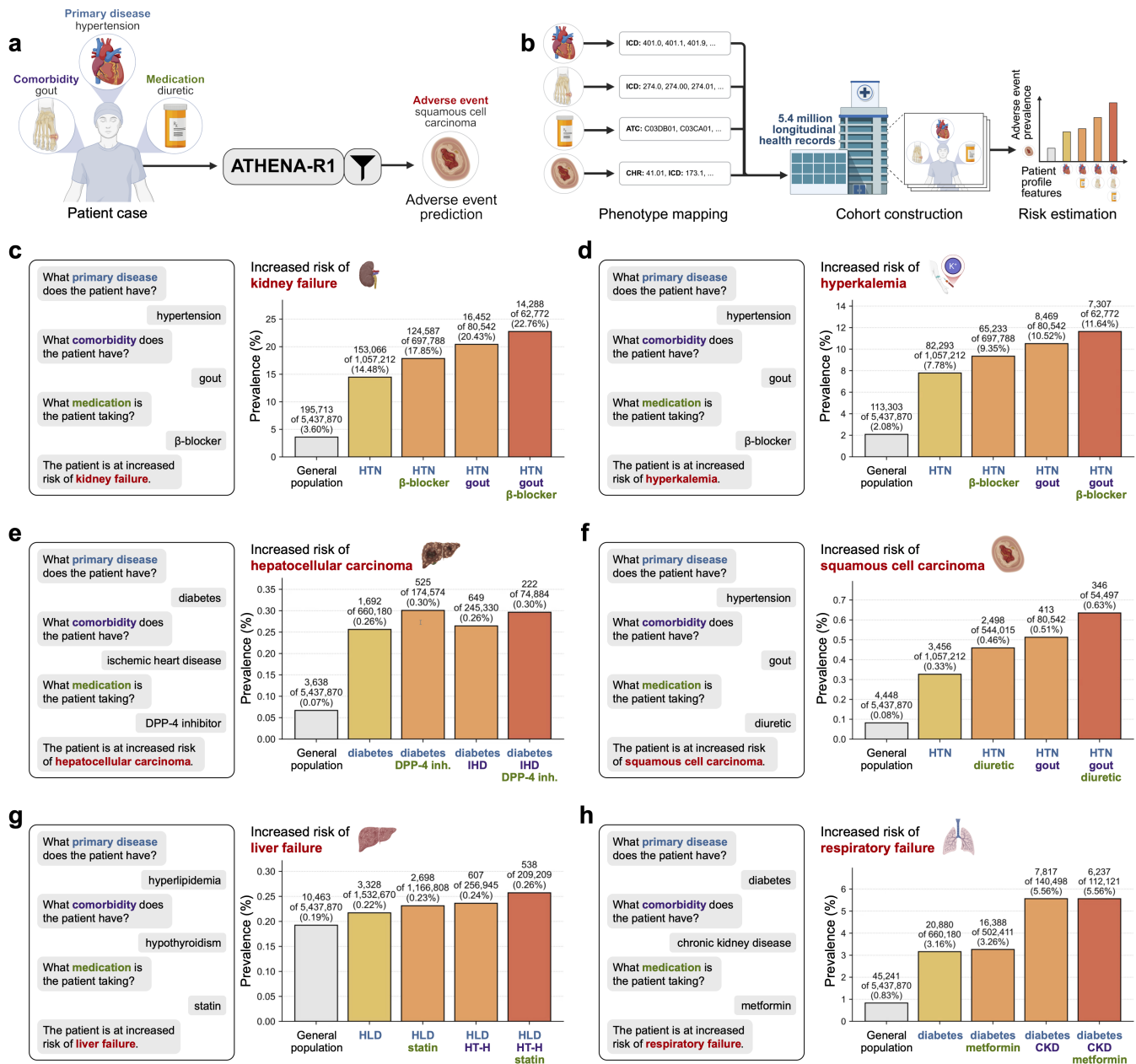


Figure 5: Adverse events predicted by ATHENA-R1 for disease, comorbidity, and medication profiles occur at the highest prevalence in the most specific patient subpopulations across electronic health records from 5.4 million patients. (a) Workflow for generating adverse-event hypotheses with ATHENA-R1. We defined patient profiles, each specified by a primary disease, a comorbidity, and a medication, and used each profile to construct a contrastive prompt that directs ATHENA-R1 to predict adverse events attributable to the full combination rather than to any single component. Candidate predictions were then scored and ranked by a clinician and a large language model. **(b)** Retrospective validation pipeline using population-scale EHRs. Clinical entities (diseases, comorbidities, drugs, and predicted adverse events) are mapped to standardized medical codes. We construct patient cohorts by identifying individuals who meet the specified clinical criteria, with index dates set by the timing of diagnoses and drug exposures. Statistical analyses then quantify the risk of each predicted adverse event in exposed versus unexposed populations, using prevalence calculations and confounder-adjusted logistic regression. **(c-h)** Evaluation of six adverse-event hypotheses generated by ATHENA-R1 across patient populations drawn from electronic health records of more than 5.4 million individuals in the Clalit Health Services. Each panel shows the prompt provided to ATHENA-R1 (left), the predicted adverse event (center), and the observed prevalence of that event across five progressively more specific patient cohorts (right). In all six panels, prevalence is highest in the most specific cohort, indicating that ATHENA-R1 identifies adverse events that are most prevalent in narrowly defined patient subpopulations. Confounder-adjusted effect estimates and control analyses for these hypotheses are shown in Figure 6.

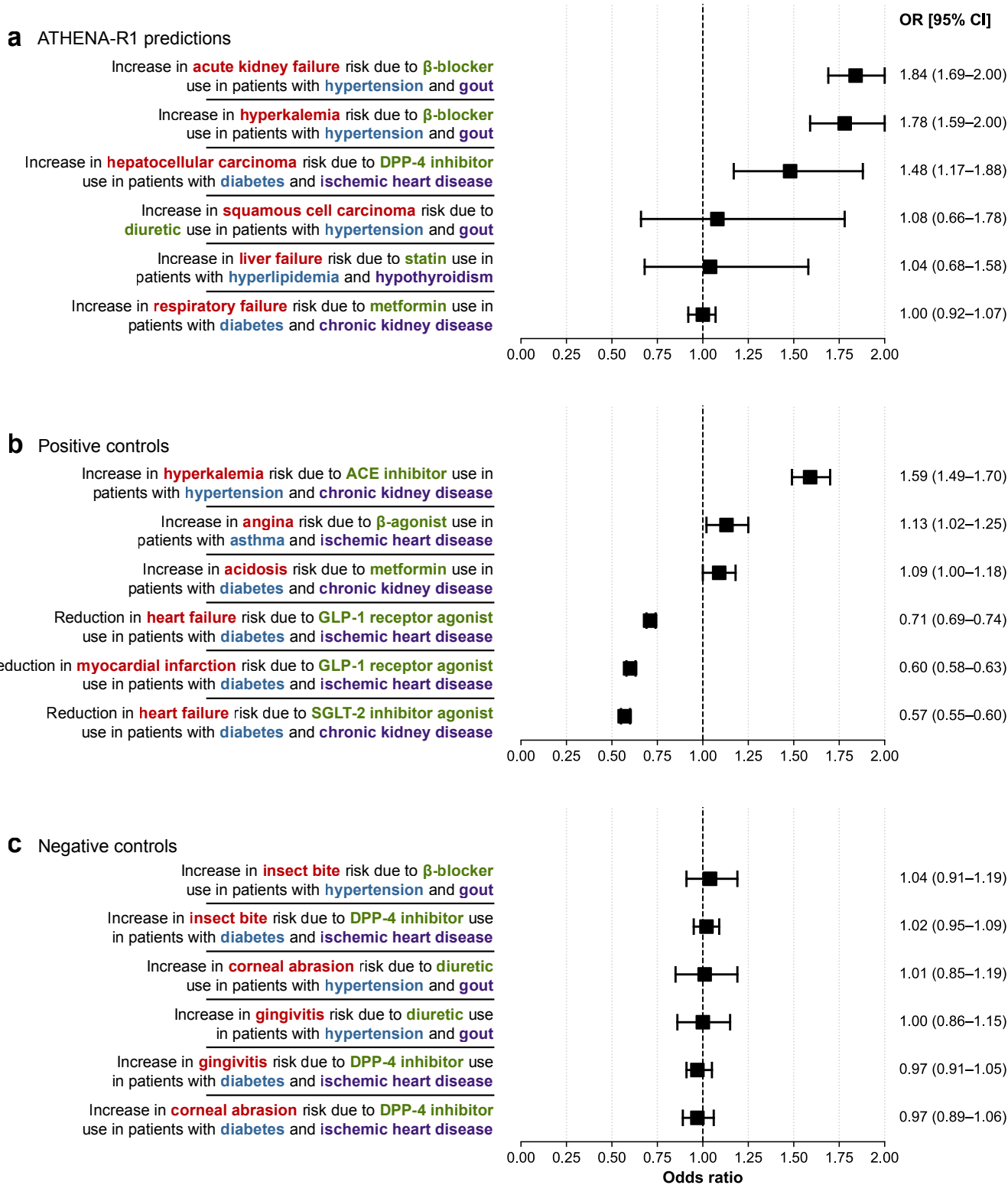


Figure 6: Population-scale electronic health records support ATHENA-R1’s adverse-event risk predictions. Forest plot of adjusted odds ratios (OR) with 95% confidence intervals (CI) for each adverse event predicted by ATHENA-R1, shown alongside positive controls and negative controls (red). Positive controls are established drug-risk associations that a calibrated pipeline should recover, and negative controls are biologically implausible associations that it should not. All regression models were adjusted for demographic confounders, including age, sex, and socioeconomic status. An OR above 1 whose CI excludes 1 denotes a statistically significant increase in risk.

(a) ATHENA-R1 predictions: adverse-event risks predicted by ATHENA-R1 for patient profiles defined by a disease, a comorbidity, and a medication. Three of six predicted associations reach statistical significance (OR > 1, 95% CI excluding 1): acute kidney failure (OR 1.84; 95% CI 1.69–2.00), hyperkalemia (OR 1.78; 95% CI 1.59–2.00), and hepatocellular carcinoma (OR 1.48; 95% CI 1.17–1.88). All six have OR ≥ 1. **(b)** Positive controls: established clinical associations used to benchmark the EHR analysis pipeline, including known risks such as hyperkalemia from ACE inhibitor use in chronic kidney disease (OR 1.59) and known protective effects such as reduced heart failure risk with SGLT-2 inhibitors (OR 0.57). **(c)** Negative controls: associations between the target drugs and unrelated medical conditions. In all cases, OR = 1, with confidence intervals crossing the vertical line of no effect.

References

1. Hamburg, M. A. & Collins, F. S. The path to personalized medicine. *New England Journal of Medicine* **363**, 301–304 (2010).
2. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* **25**, 44–56. doi:[10.1038/s41591-018-0300-7](https://doi.org/10.1038/s41591-018-0300-7) (2019).
3. Dubey, A. *et al.* The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
4. Singhal, K. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
5. Singhal, K. *et al.* Toward expert-level medical question answering with large language models. *Nature Medicine*, 1–8 (2025).
6. Chen, Z. *et al.* Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079* (2023).
7. McDuff, D. *et al.* Towards accurate differential diagnosis with large language models. *Nature* **642**, 451–457 (2025).
8. Gao, S. *et al.* Empowering biomedical discovery with AI agents. *Cell* **187**, 6125–6151. doi:[10.1016/j.cell.2024.09.022](https://doi.org/10.1016/j.cell.2024.09.022) (2024).
9. Tu, T. *et al.* Towards conversational diagnostic artificial intelligence. *Nature* **642**, 442–450 (2025).
10. Gao, Y. *et al.* Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* (2023).
11. Yan, F. *et al.* Berkeley Function Calling Leaderboard https://gorilla.cs.berkeley.edu/blogs/8_berkeley_function_calling_leaderboard.html. 2024.
12. Dadao, I. *watt-tool-8B: A Fine-Tuned Language Model for Tool Usage and Multi-Turn Dialogue* <https://huggingface.co/watt-ai/watt-tool-8B>. 2025.
13. Liu, W. *et al.* ToolACE: Winning the Points of LLM Function Calling in The Thirteenth International Conference on Learning Representations (ICLR) (2025).
14. Singh, A. *et al.* OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267* (2025).
15. Guo, D. *et al.* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **645**, 633–638. doi:[10.1038/s41586-025-09422-z](https://doi.org/10.1038/s41586-025-09422-z) (2025).
16. Kass-Hout, T. A. *et al.* OpenFDA: an innovative platform providing access to a wealth of FDA’s publicly available data. *Journal of the American Medical Informatics Association* **23**, 596–600 (2016).

17. Ochoa, D. *et al.* The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research* **51**, D1353–D1359. doi:[10.1093/nar/gkac1046](https://doi.org/10.1093/nar/gkac1046) (2023).
18. Yao, S. *et al.* *React: Synergizing reasoning and acting in language models* in *International Conference on Learning Representations (ICLR)* (2023).
19. Zelikman, E., Wu, Y., Mu, J. & Goodman, N. D. *STaR: Bootstrapping Reasoning With Reasoning* in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022 (NeurIPS 2022)* (eds Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. & Oh, A.) (2022).
20. Wang, Y. *et al.* *Self-Instruct: Aligning Language Models with Self-Generated Instructions* in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (eds Rogers, A., Boyd-Graber, J. & Okazaki, N.) (Association for Computational Linguistics, Toronto, Canada, 2023), 13484–13508. doi:[10.18653/v1/2023.acl-long.754](https://doi.org/10.18653/v1/2023.acl-long.754).
21. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35**, 27730–27744 (2022).
22. Qian, C. *et al.* *ToolRL: Reward is All Tool Learning Needs* in *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025 (NeurIPS 2025)* (2025).
23. Shao, Z. *et al.* Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* (2024).
24. Zhao, Y. *et al.* *Geometric-mean policy optimization* in *International Conference on Learning Representations (ICLR)* (2026).
25. Golchin, S. & Surdeanu, M. *Time Travel in LLMs: Tracing Data Contamination in Large Language Models* in *Proceedings of the 12th International Conference on Learning Representations (ICLR 2024)* (2024).
26. Su, X. *et al.* *KGAREvion: An AI Agent for Knowledge-Intensive Biomedical QA* in *Proceedings of the 13th International Conference on Learning Representations (ICLR 2025)* (2025).
27. Huang, K. *et al.* A foundation model for clinician-centered drug repurposing. *Nature Medicine* **30**, 3601–3613 (2024).
28. Bristol-Myers Squibb Company. *KENALOG-10 Injection (triamcinolone acetonide injectable suspension, USP): Prescribing Information* 2018.
29. Ahmet, A. *et al.* Adrenal suppression from glucocorticoids: preventing an iatrogenic cause of morbidity and mortality in children. *BMJ Paediatrics Open* **3**, e000569. doi:[10.1136/bmjpo-2019-000569](https://doi.org/10.1136/bmjpo-2019-000569) (2019).
30. Tambuyzer, E. *et al.* Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nature Reviews Drug Discovery* **19**, 93–111 (2020).
31. Boycott, K. M. & Ardigó, D. Addressing challenges in the diagnosis and treatment of rare genetic diseases. *Nature Reviews Drug Discovery* **17**, 151–152 (2018).

32. Chiang, W.-L. *et al.* Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference in *Proceedings of the 41st International Conference on Machine Learning* (eds Salakhutdinov, R. *et al.*) **235** (PMLR, 2024), 8359–8388.
33. Pfohl, S. R. *et al.* A toolbox for surfacing health equity harms and biases in large language models. *Nature Medicine* **30**, 3590–3600 (2024).
34. Heidenreich, P. A. *et al.* 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. *Circulation* **145**, e895–e1032. doi:[10.1161/CIR.0000000000001063](https://doi.org/10.1161/CIR.0000000000001063) (2022).
35. Holbrook, A. M. *et al.* Systematic overview of warfarin and its drug and food interactions. *Archives of Internal Medicine* **165**, 1095–1106. doi:[10.1001/archinte.165.10.1095](https://doi.org/10.1001/archinte.165.10.1095) (2005).
36. Salpeter, S. R., Ormiston, T. M. & Salpeter, E. E. Cardioselective β -Blockers in Patients with Reactive Airway Disease: A Meta-Analysis. *Annals of Internal Medicine* **137**, 715–725. doi:[10.7326/0003-4819-137-9-200211050-00035](https://doi.org/10.7326/0003-4819-137-9-200211050-00035) (2002).
37. Zheng, L. *et al.* Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023 (NeurIPS 2023) Datasets and Benchmarks Track* (2023).
38. Sherman, R. E. *et al.* Real-World Evidence – What Is It and What Can It Tell Us? *New England Journal of Medicine* **375**, 2293–2297. doi:[10.1056/NEJMs1609216](https://doi.org/10.1056/NEJMs1609216) (2016).
39. Hernán, M. A. & Robins, J. M. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *American Journal of Epidemiology* **183**, 758–764. doi:[10.1093/aje/kwv254](https://doi.org/10.1093/aje/kwv254) (2016).
40. Albasri, A. *et al.* Association between antihypertensive treatment and adverse events: systematic review and meta-analysis. *BMJ* **372**, n189. doi:[10.1136/bmj.n189](https://doi.org/10.1136/bmj.n189) (2021).
41. Palmer, B. F. Renal Dysfunction Complicating the Treatment of Hypertension. *New England Journal of Medicine* **347**, 1256–1261. doi:[10.1056/NEJMra020676](https://doi.org/10.1056/NEJMra020676) (2002).
42. Choi, H. K., Soriano, L. C., Zhang, Y. & Rodriguez, L. A. G. Antihypertensive drugs and risk of incident gout among patients with hypertension: population based case-control study. *BMJ* **344**, d8190–d8190. doi:[10.1136/bmj.d8190](https://doi.org/10.1136/bmj.d8190) (2012).
43. Xu, X., Hu, J., Song, N., Chen, R., Zhang, T. & Ding, X. Hyperuricemia increases the risk of acute kidney injury: a systematic review and meta-analysis. *BMC Nephrology* **18**, 27. doi:[10.1186/s12882-016-0433-1](https://doi.org/10.1186/s12882-016-0433-1) (2017).
44. Chang, A. R. *et al.* Antihypertensive Medications and the Prevalence of Hyperkalemia in a Large Health System. *Hypertension* **67**, 1181–1188. doi:[10.1161/HYPERTENSIONAHA.116.07363](https://doi.org/10.1161/HYPERTENSIONAHA.116.07363) (2016).
45. Rosa, R. M. *et al.* Adrenergic Modulation of Extrarenal Potassium Disposal. *New England Journal of Medicine* **302**, 431–434. doi:[10.1056/NEJM198002213020803](https://doi.org/10.1056/NEJM198002213020803) (1980).
46. Li, X., Wang, X. & Gao, P. Diabetes Mellitus and Risk of Hepatocellular Carcinoma. *BioMed Research International* **2017**, 1–10. doi:[10.1155/2017/5202684](https://doi.org/10.1155/2017/5202684) (2017).

47. Do, D., Lee, T., Peasah, S., Inneh, A., Patel, U. & Good, C. Trends in first-line glucose-lowering medication use among US adults with type 2 diabetes from 2019 to 2023. *Journal of Managed Care & Specialty Pharmacy* **31**, 520–526. doi:[10.18553/jmcp.2025.31.5.520](https://doi.org/10.18553/jmcp.2025.31.5.520) (2025).
48. Zhao, M. *et al.* Dipeptidyl peptidase-4 inhibitors and cancer risk in patients with type 2 diabetes: a meta-analysis of randomized clinical trials. *Scientific Reports* **7**, 8273. doi:[10.1038/s41598-017-07921-2](https://doi.org/10.1038/s41598-017-07921-2) (2017).
49. Yang, J., Hwang, Y., Ju, J.-S., Han, S., An, J. & Shim, J. H. Impact of newer antihyperglycemic agents on hepatic complications: A systematic review and meta-analysis of data from 5.3 million patients with type 2 diabetes mellitus. *Hepatology*. doi:[10.1097/HEP.0000000000001695](https://doi.org/10.1097/HEP.0000000000001695) (2026).
50. Kawakita, E., Koya, D. & Kanasaki, K. CD26/DPP-4: Type 2 Diabetes Drug Target with Potential Influence on Cancer Biology. *Cancers* **13**, 2191. doi:[10.3390/cancers13092191](https://doi.org/10.3390/cancers13092191) (2021).
51. Schneider, R., Reinau, D., Stoffel, S., Jick, S., Meier, C. & Spoendlin, J. Risk of skin cancer in new users of thiazides and thiazide-like diuretics: a cohort study using an active comparator group*. *British Journal of Dermatology* **185**, 343–352. doi:[10.1111/bjd.19880](https://doi.org/10.1111/bjd.19880) (2021).
52. Tian, L., Wang, Y., Zhang, Y., Tian, L. & Wang, H. Association between gout and cancers: A systematic review and meta-analysis. *Medicine* **103**, e40234. doi:[10.1097/MD.00000000000040234](https://doi.org/10.1097/MD.00000000000040234) (2024).
53. Newman, C. B. *et al.* Statin Safety and Associated Adverse Events: A Scientific Statement From the American Heart Association. *Arteriosclerosis, Thrombosis, and Vascular Biology* **39**. doi:[10.1161/ATV.0000000000000073](https://doi.org/10.1161/ATV.0000000000000073) (2019).
54. Yen, F.-S., Wei, J. C.-C., Yang, Y.-C., Hsu, C.-C. & Hwu, C.-M. Respiratory outcomes of metformin use in patients with type 2 diabetes and chronic obstructive pulmonary disease. *Scientific Reports* **10**, 10298. doi:[10.1038/s41598-020-67338-2](https://doi.org/10.1038/s41598-020-67338-2) (2020).
55. Weinberg, J. M. *et al.* Risk of hyperkalemia in nondiabetic patients with chronic kidney disease receiving antihypertensive therapy. *Archives of Internal Medicine* **169**, 1587–1594. doi:[10.1001/archinternmed.2009.284](https://doi.org/10.1001/archinternmed.2009.284) (2009).
56. Au, D. H., Curtis, J. R., Every, N. R., McDonnell, M. B. & Fihn, S. D. Association between inhaled beta-agonists and the risk of unstable angina and myocardial infarction. *Chest* **121**, 846–851. doi:[10.1378/chest.121.3.846](https://doi.org/10.1378/chest.121.3.846) (2002).
57. DeFronzo, R., Fleming, G. A., Chen, K. & Bicsak, T. A. Metformin-associated lactic acidosis: Current perspectives on causes and risk. *Metabolism: Clinical and Experimental* **65**, 20–29. doi:[10.1016/j.metabol.2015.10.014](https://doi.org/10.1016/j.metabol.2015.10.014) (2016).
58. McGuire, D. K. *et al.* Oral Semaglutide and Cardiovascular Outcomes in High-Risk Type 2 Diabetes. *The New England Journal of Medicine* **392**, 2001–2012. doi:[10.1056/NEJMoa2501006](https://doi.org/10.1056/NEJMoa2501006) (2025).
59. Pop-Busui, R. *et al.* Oral Semaglutide and Heart Failure Outcomes in Persons With Type 2 Diabetes: A Secondary Analysis of the SOUL Randomized Clinical Trial. *JAMA internal medicine* **186**, 426–436. doi:[10.1001/jamainternmed.2025.7774](https://doi.org/10.1001/jamainternmed.2025.7774) (2026).

60. Natale, P. *et al.* Sodium-glucose co-transporter protein 2 (SGLT2) inhibitors for people with chronic kidney disease and diabetes. *The Cochrane Database of Systematic Reviews* **5**, CD015588. doi:[10.1002/14651858.CD015588.pub2](https://doi.org/10.1002/14651858.CD015588.pub2) (2024).
61. Hager, P. *et al.* Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine* **30**, 2613–2622 (2024).
62. Ghassemi, M., Oakden-Rayner, L. & Beam, A. L. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* **3**, e745–e750. doi:[10.1016/S2589-7500\(21\)00208-9](https://doi.org/10.1016/S2589-7500(21)00208-9) (2021).
63. Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R. & Gal, Y. AI models collapse when trained on recursively generated data. *Nature* **631**, 755–759 (2024).
64. Moor, M. *et al.* Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265. doi:[10.1038/s41586-023-05881-4](https://doi.org/10.1038/s41586-023-05881-4) (2023).